# Estimating the prevalence of infectious diseases from under-reported age-dependent compulsorily notification databases

***Marcos Amaku[1], Marcelo Nascimento Burattini[1,2], Eleazar Chaib,[1] Francisco Antonio Bezerra Coutinho[1], David Greenhalgh[3], Luis Fernandez Lopez[1,4] and Eduardo Massad[1,5,*]***

[1] LIM01-Hospital de Clínicas, Faculdade de Medicina Universidade de São Paulo, São Paulo, SP, Brazil

[2] Hospital São Paulo, Escola Paulista de Medicina, Universidade Federal de São Paulo, São Paulo, SP, Brazil

[3] Department of Mathematics and Statistics, The University of Strathclyde, Glasgow, Scotland, UK

[4] Center for Internet Augmented Research & Assessment, Florida International University, Miami, FL, USA

[5] London School of Hygiene and Tropical Medicine, London, UK

*corresponding author

Authors listed in alphabetical order

E-mail addresses:

MA: amaku@usp.br

MNB: mnburatt@gmail.com

EC: eleazarchaib@yahoo.co.uk

FABC: coutinho@dim.fm.usp.br

DG: david.greenhalgh@strath.ac.uk

LFL: lopez@wfpd.com.br

EM: edmassad@dim.fm.usp.br

**Abstract**

**Background:** National or local laws, norms or regulations (sometimes and in some countries) require medical providers to report notifiable diseases to public health authorities. Reporting, however, is almost always incomplete. This is due to a variety of reasons, ranging from not recognizing the diseased to failures in the technical or administrative steps leading to the final official register in the disease notification system. The reported fraction varies from 9% to 99% and is strongly associated with the disease being reported.

**Methods:** In this paper we propose a method to approximately estimate the full prevalence (and any other variable or parameter related to transmission intensity) of infectious diseases. The model assumes incomplete notification of incidence and allows the estimation of the non-notified number of infections and it is illustrated by the case of hepatitis C in Brazil. The method has the advantage that it can be corrected iteratively by comparing its findings with empirical results.

**Results:** The application of the model for the case of hepatitis C in Brazil resulted in a prevalence of notified cases that varied between 163,902 and 169,382 cases; a prevalence of non-notified cases that varied between 1,433,638 and 1,446,771; and a total prevalence of infections that varied between 1,597,540 and 1,616,153 cases.

**Conclusions:** We conclude that that the model proposed can be useful for estimation of the actual magnitude of endemic states of infectious diseases, particularly for those where the number of notified cases is only the tip of the iceberg. In addition, the method can be applied to other situations, such as the well-known underreported incidence of criminality (for example rape), among others.

**Keywords:** Hepatitis C; Mathematical Models; Notifications System Incidence; Prevalence.

**Background**

Compulsory notifiable diseases (CNDs) are those diseases that should be compulsorily reported to Health Authorities as soon as suspected by the attending professional [1]. The notified cases then enter a database from which, among other things, it is possible to know the incidence (new cases per age, sex, risk factor, geographic location, etc, per period of time) of the disease. The availability of such information allows health authorities, in principle, to monitor and to plan controlling the disease, for example providing early warning of possible outbreaks [2].

In spite of international, national or local laws, norms or regulations requiring medical providers to report notifiable diseases to public health authorities, reporting is almost always incomplete [3-8]. This is due to a variety of reasons. First diseases may be asymptomatic. For example only around one in five dengue cases are symptomatic [9]. Second a case may be symptomatic but an individual may not seek healthcare due to mild or self-limiting symptoms or lack of knowledge about when to seek healthcare [4] or social stigma due to the nature of the disease, (for example sexually transmitted diseases). Even if an individual seeks healthcare a disease may not be notifiable, or if now notifiable may not have been notifiable in the past leading to incomplete notification records. A disease may also be misdiagnosed. Finally there may be failures in the technical or administrative steps leading to registration [10].

Rosenburg et al. [11] estimated that for every 100 persons infected with *Shigella*, 76 become symptomatic, 28 consulted a physician, nine submitted stool samples, seven had positive results, six were reported to the local health department and five were reported nationally to the Centers for Disease Control and Prevention. Thus they proposed a multiplication factor of 20 to estimate the number of *Shigella* infections based on national *Shigellosis* case reports.

Konowitz, Petrossian and Rose [10] investigated under-reporting of disease and knowledge of physicians of reporting requirements at two hospitals in New York City in 1982. They say that physicians may not know which diseases are reportable or the

correct reporting procedures. The percentage of physicians who knew which diseases they had to report ranged from 37% for trachoma to 96% for syphilis. The results of Konowitz et al. suggest that a major factor in physician under-reporting is lack of knowledge of the reporting system.

Brabazon et al. [12] highlighted the extent of under-reporting of notifiable infectious disease hospitalisations in a health-board in Ireland, which was felt to be concerning for disease surveillance. Under-reporting was definitely demonstrated in 9 out of 22 notifiable diseases amounting to 572 cases (18% of missed cases). The most missed cases were viral meningitis, infectious mononucleosis, unspecified hepatitis C and acute encephalitis.

Keramou and Evans [5] performed a systematic review of completeness of infectious disease notification in the United Kingdom. Reporting completeness varied from 3% to 95% and was most strongly correlated with the disease being reported. Median reporting completeness was 73% for tuberculosis, 65% for meningococcus disease and 40% for other diseases. They conclude that reporting completeness remains suboptimal even for diseases that are under enhanced surveillance or were of significant public health importance.

A review by Doyle et al. [3], limited to published studies conducted in the United States between 1970 and 1999, quantitatively assessed infectious disease reporting completeness and found that reporting completeness varied from 9% to 99% and was strongly associated with the disease being reported. In another study [13] the mean reporting completeness for acquired immunodeficiency syndrome, sexually transmitted diseases, and tuberculosis as a group was significantly higher (79%) than for all other diseases combined (49%).

Schiffman et al. [14] investigated under-reporting of lyme and other tick-borne diseases in residents of a high incidence county, Minnesota, USA, in 2009. Of 444 illness events 352 (79%) were not reported. Of these 102 (29%) meet confirmed or probable surveillance case criteria including 91 (26%) confirmed lyme disease cases.

Serra et al. [8] developed a universal method to correct under-reporting of communicable diseases and applied it to incidence of hydatidosis in Chile, 1985-1994. According to this method the real rate of human hydatidosis in the period 1985-1994 was four times higher than the official notification in the given period.

Rowe and Cowie [6] used data linkage to improve the completeness of Aboriginal and Torres Strait Islander status in communicable disease notifications in Victoria, Australia. The burden of notifiable diseases in Torres Strait Islander Victorians could not be accurately estimated due to under-reporting of indigenous status. There were 12,488 cases of hepatitis B, hepatitis C (HCV) and gonococcal infection in Victoria in 2009-2010 with indigenous status missing in 61.6%, 67.8% and 33.1% of those conditions, respectively. They used data linkage to improve completeness of indigenous status in people notified with viral hepatitis and gonococcal infection.

Of particular concern are those chronic, mainly asymptomatic, infectious diseases that allow infected individuals to live for years or even decades without being recognised as such. These diseases can represent a heavy burden to the affected populations and pose significant risk to the international community. Perhaps the most dramatic examples of the latter include human immunodeficiency (HIV) and HCV viruses pandemics. In fact, these two infections have been labeled by WHO as the epidemics of the XXth and XXIth centuries, respectively [7, 15].

One critical consequence of under-notification of such diseases is the fact that their prevalence estimates are frequently way under-estimated, leading to miscalculation of their actual burden and making control efforts suboptimal [4].

HCV is a disease with a long period between infection and symptoms developing. Because infected people are mainly asymptomatic and risk behaviour may have occurred a long time ago individuals often do not consult health professionals to discuss potential disease infection. As in general a large high risk group is people who share injection equipment and other injection paraphernalia, for example cookers, filters and spoons, and drug injection is an illegal activity, which often does not meet with social approval, light to moderate injectors, or past injectors who do not currently inject, may not disclose their risky behaviour to their health provider. Being unaware of the risk

behaviour the health provider is unable to recommend HCV screening. Also HCV is extremely easy to catch via injecting. Past injectors who no longer inject may not perceive themselves to be at risk.

In a previous paper [16] we assumed that the infection (HCV) was in steady-state. Then we proposed two methods to give a first rough estimate of the actual number of HCV infected individuals (prevalence) taking into account the yearly notification rate of newly reported infections (incidence of notification) and the size of the Liver Transplantation Waiting List (LTWL) of patients with liver failure due to chronic HCV infection [17]. Both approaches, when applied to the Brazilian HCV situation converged to the same results, that is, the methods proposed reproduce both the prevalence of reported cases and the LTWL with reasonable accuracy. In that paper we show how to calculate the prevalence of people living with HCV in Brazil, which resulted in a value up to 8 times higher than the official reported number of cases [16].

In both [16] and this paper the under-reporting mechanism is included in the model by dividing the infected individuals into two categories: notified and non-notified. Newly infected individuals enter the non-notified class and leave it either through death, recovery or notification. If they are notified they immediately enter the notified infected class.

The present paper is an improvement of those techniques because, unlike in the previous paper mentioned above, now we do not assume steady state. Unfortunately, given the short period of time with data available (hepatitis notification became compulsory in Brazil only in 1999 [18], it cannot give more precise information on HCV prevalence than the one already provided by our previous study, but it illustrates the techniques that allow the prevalence estimation based on age and time of previous notifications, and that can be applied to any notifiable disease.

This paper is organised as follows: First we describe a continuous model, that is a model where the variables are continuous functions of age and time. Next we describe a discrete model, in which the variables are discrete functions of age and time. In the following section we discuss application to HCV. Then we turn to our estimation

method applied to the size of the Liver Transplantation Waiting List in Brazil. The next section gives our numerical results. Discussion and conclusions close the paper.

## Methods

### Continuous time and age model

Assume we have an SIR (Susceptible-Infected-Removed) type infection and let $S(a,t)da$, $I(a,t)da$ and $R(a,t)da$ be the number of individuals with age between $a$ and $a+da$ at time $t$ that are susceptible, infected and removed (or recovered), respectively. In addition, as mentioned in the Background section, public health authorities demand that some diseases be compulsorily notifiable, that is they publish the number of diagnosed individuals per time unit for each age interval (incidence) in public databases. Therefore, we can divide the prevalence of infected individuals into two classes: notified individuals, denoted $I^N(a,t)da$, and non-notified individuals, denoted $I^{NN}(a,t)da$.

Let $\lambda(a,t)$ be the so-called age and time-dependent force-of-infection (incidence density). Then:

$$\lambda(a,t)S(a,t)dadt \tag{1}$$

is the number of susceptible individuals who get the infection when aged between $a$ and $a+da$ during the time interval $dt$. Standard arguments allow us to write the following system of partial differential equations, known as Trucco-Von Foester equations in the literature [19] :

$$\begin{aligned}
&\frac{\partial S(a,t)}{\partial t}+\frac{\partial S(a,t)}{\partial a}=-\lambda(a,t)S(a,t)-\mu(a,t)S(a,t),\\
&\frac{\partial I^{NN}(a,t)}{\partial t}+\frac{\partial I^{NN}(a,t)}{\partial a}=\lambda(a,t)S(a,t)\\
&\qquad -\left(\mu(a,t)+\alpha^{NN}(a,t)+\gamma^{NN}(a,t)\right)I^{NN}(a,t)-\kappa(a,t)I^{NN}(a,t),\\
&\frac{\partial I^{N}(a,t)}{\partial t}+\frac{\partial I^{N}(a,t)}{\partial a}=\kappa(a,t)I^{NN}(a,t)-\left(\mu(a,t)+\alpha^{N}(a,t)+\gamma^{N}(a,t)\right)I^{N}(a,t),
\end{aligned} \tag{2}$$

$$\frac{\partial R(a,t)}{\partial t} + \frac{\partial R(a,t)}{\partial a} = \gamma^{NN}(a,t)I^{NN}(a,t) + \gamma^{N}(a,t)I^{N}(a,t) - \mu(a,t)R(a,t),$$

where the meaning of the parameters is described in Table 1.

**Table 1. Parameters used in model (2).**

| Parameter | Meaning | Values used in the numerical simulation |
|---|---|---|
| $\lambda(a,t)$ | Force of Infection | Calculated |
| $\mu(a,t)$ | Natural Mortality Rate | 0.0133 year$^{-1}$* |
| $\alpha^{NN}(a,t)$ | Disease-induced Mortality Rate for non-notified individuals | ** |
| $\alpha^{N}(a,t)$ | Disease-induced Mortality Rate for notified individuals | ** |
| $\gamma^{NN}(a,t)$ | Recovery Rate for non-notified individuals | Assumed negligible |
| $\gamma^{N}(a,t)$ | Recovery Rate for notified individuals | Assumed negligible |
| $\kappa(a,t)$ | Notification Rate | 0.0125 [16] |

* From demographic data of Brazil.

** Constructed as equal to $0.15/\{1 + \exp(-0.1(a - 57.31))\}$ years$^{-1}$ as in [16].

In Table 1, we neglected the value of the recovery rates in the numerical simulations because we assumed that HCV infection is very long-lasted. These parameters, however, were included in the model for the sake of completeness.

The notification rate $\kappa(a,t)$ is one of the most important parameters in the model. This represents the rate at which those non-notified individuals of age *a* are reported to health authorities and notified. This has two components, first the rate of an infected person being recognised and secondly the rate of being reported. So if $\kappa(a,t)$ is small then there will be a large number of non-notified infected individuals hidden from the system, whereas if $\kappa(a,t)$ is large then most infected individuals will be notified and the records will accurately reflect the number infected in the population.

The solution of system (2) can be obtained with the method of characteristics [19]. However, for our purposes, it is better to solve the equation by following a cohort, as described in [20].

The solution of the equation for susceptible individuals is:

$$S(a, t_0 + a) = S(0, t_0)\exp\left(-\int_0^a [\lambda(s, t_0 + s) + \mu(s, t_0 + s)]ds\right). \quad (3)$$

There are a small number of maternal-infant HCV infections [21]. It would be possible to include these in the theoretical model. However data for age zero is not used in the calculations because it is unreliable. So to include maternal-infant HCV infections would make the model more complicated but not change the numerical results. So we ignore these maternal-infant HCV infections.

The solution for the equation for infected individuals is:

$$I^{NN}(a, t_0 + a) = \int_0^a \lambda(s, t_0 + s)\, S(s, t_0 + s)$$
$$\exp\left(-\int_s^a [\mu(x, t_0 + x) + \gamma^{NN}(x, t_0 + x) + \alpha^{NN}(x, t_0 + x) + \kappa(x, t_0 + x)]dx\right) ds, \quad (4)$$
$$I^{N}(a, t_0 + a) = \int_0^a \kappa(s, t_0 + s)\, I^{NN}(s, t_0 + s)$$
$$\exp\left(-\int_s^a [\mu(x, t_0 + x) + \gamma^{N}(x, t_0 + x) + \alpha^{N}(x, t_0 + x)]dx\right) ds. \quad (5)$$

Finally, the equation for the removed individuals is given by:

$$R(a, t_0 + a) = \int_0^a \left(\gamma^{NN}(s, t_0 + s) I^{NN}(s, t_0 + s) + \gamma^{N}(s, t_0 + s) I^{N}(s, t_0 + s)\right)$$
$$\exp\left(-\int_s^a [\mu(x, t_0 + x)]dx\right) ds. \quad (6)$$

Assuming steady state, the system (1) was solved by Amaku et al. [16] to calculate the prevalence of HCV in Brazil. The work that follows is an extension of the methods described there and its results are in accordance with the previous results for the cases where real data are available.

**Discrete time and age model**

In real life epidemics notification is discrete with the time and age units expressed in weeks, months or years. Hence, in order to apply the model to a real public health problem we discretised model (2), with time and age unit expressed in years. This

discretisation has to be done carefully to use the maximum advantage of the data available.

**Calculating the prevalence $I^{NN*}\{A,i\}$ and $I^{N*}\{A,i\}$**

To avoid potential confusion between similar variables in the discrete and continuous models we adopt the convention that discrete variables have a '*' superscript after the variable and their arguments are in curly parentheses,{ }, whereas continuous variables do not have a '*' superscript after the variable and their arguments are in round parentheses ().

From the SINAN database we can calculate $SINAN*\{A,i\}$ where $A$ is an integer number and $i$ represents a calendar year, which represents the number of infected individuals notified to SINAN in the calendar year $i$, who at the end of calendar year $i$ have age $A$ years (in other words at the end of calendar year $i$ their exact age $a$ is in the time interval $[A,A+1)$).

Because we want the variables in the discrete model to relate to the SINAN data we similarly define $I^{NN*}\{A,i\}$ and $I^{N*}\{A,i\}$ to denote respectively the number of non-notified infected and notified infected individuals at time the end of calendar year $i$, whose age at that time is $A$ years (so their exact age lies in $[A,A+1)$). Given parametric functions such as $\kappa(a,t)$ and $\phi^{NN}(a,t)$ in the continuous model, in the corresponding discrete model these are assumed to be discrete functions $\kappa_d(a,t) = \kappa_{A,i}$ and $\phi_d^{NN}(a,t) = \phi_{A,i}^{NN}$ for $(a,t) \in R=\{a \in [A, A+1)$ and $t \in (t_i\text{-}1, t_i]\}$. Here $t_i$ denotes the end of calendar year $i$, and $\kappa_{A,i}$ and $\phi_{A,i}^{NN}$ are respectively the average values of $\kappa(a,t)$ and $\phi^{NN}(a,t)$ over the region $R$.

The discretised versions of equations (4) and (5) are given by equations (7) and (8) below, which are approximations as explained in the Appendix.

$$I^{NN*}\{A,i\} = I^{NN*}\{A-1,i-1\}\exp\left[-\frac{1}{2}\left(\kappa_{A-1,i} + \kappa_{A,i} + \phi_{A-1,i}^{NN} + \phi_{A,i}^{NN}\right)\right] + INC\{A,i\}, \quad (7)$$

where for $A$=0, $I^{NN*}\{A-1,i-1\} = 0$. $INC\{A,i\}$ is the new HCV cases occurring between times $t_i$-1 and $t_i$ that are still alive, infectious and non-notified at time $t_i$ in the

year cohort born between times $t_i$-$A$-1 and $t_i$-$A$. Here (using the continuous model notation)

$$\phi^{NN}(a,t) = \mu(a,t) + \gamma^{NN}(a,t) + \alpha^{NN}(a,t).$$

In equation (7), the term

$$\exp\left[-\frac{1}{2}\left(\kappa_{A-1,i} + \kappa_{A,i} + \phi^{NN}_{A-1,i} + \phi^{NN}_{A,i}\right)\right]$$

means the probability of not being removed from the non-notified class of individuals, either by natural death, disease-induced death, recovery or notification in the interval ($t_i$-1,$t_i$]. Equation (7) is very important because, as shown later in the paper, it allows the calculation of the true incidence from empirical data (see equation (12) below).

Recurrence equation (7) can be solved by well-known methods and the prevalence of notified and non-notified individuals can be estimated (see equations (13) and (14) below).

Similarly, we can write:

$$I^{N*}\{A,i\} = I^{N*}\{A-1,i-1\}\exp\left[-\frac{1}{2}\left(\phi^{N}_{A-1,i} + \phi^{N}_{A,i}\right)\right] + \int_A^{A+1} NOTIFICATION(a,(t_i-1,t_i])da, \quad (8)$$

where (again using the continuous model notation) $\phi^N(a,t) = \mu(a,t) + \gamma^N(a,t) + \alpha^N(a,t)$. The last term represents the notifications of HCV between times $t_i$-1 and $t_i$ of individuals in the year cohort born in $t_i$-$A$-1 to $t_i$-$A$ who are still in the notified class at time $t_i$, i.e.

$$\int_A^{A+1}\int_0^1 \kappa_d(a-1+x,t_i-1+x)I^{NN}(a-1+x,t_i-1+x) \exp\left[-\int_x^1 \phi^N_d(a-1+z,t_i-1+z)\,dz\right]dx\,da, \approx \kappa_d\left(A+\tfrac{1}{2},t_i\right)I^{NN}\left(A+\tfrac{1}{2},t_i\right). \quad (9)$$

This is because both integration intervals are of length one, hence to first order we can approximate the integrand by its value at any specific point in the integrated area. So we choose $a = A + \frac{1}{2}$, $x$=1. Now note that

(i) $\kappa_d\left(A + \frac{1}{2}, t_i\right) = \kappa_{A,i}$, as in the discrete model $\kappa_d(a,t) = \kappa_{A,i}$ over the region

$R$={$a \in [A, A+1)$ and $t \in (t_i\text{-}1, t_i]$},

and

(ii) $I^{NN*}\{A,i\} \approx I^{NN}\left(A + \frac{1}{2}, t_i\right)$,

as explained in the Appendix (equation (A5)). Hence the last term in (8) is

$$\int_A^{A+1} NOTIFICATION(a,(t_i-1,t_i])da \approx \kappa_{A,i}\, I^{NN*}\{A,i\}. \qquad (10)$$

In the next section, we are going to show how to solve equations (7) and (8) using the notified cases in a particular setting, namely HCV in Brazil. Using the notified incidences and good guesses for the mortality rates we can calculate any desired properties of the infected population. In the next section we calculate the prevalence of the disease. The calculation presented applies to any notifiable infectious disease.

**Example of Application: Hepatitis C**

In this section we exemplify the above theory by calculating the prevalence of HCV, a flaviviral infection that afflicts close to 3% of the world population [22], in Brazil. As mentioned in the Introduction, the great majority of infections with HCV, however, are not easily identified and, therefore, frequently non-notified. Our data were taken from the National Reportable Disease Information System "Sistema de Informação de Agravos de Notificação" (SINAN) of the Brazilian Health Ministry [23]. SINAN is publicly available through the internet and used by the World Health Organisation [24]. It is used throughout Brazil, in all health institutions whether public or private. All Brazilians diagnosed with HCV are reported to SINAN. The database includes symptomatic patients who report to a doctor, also symptomatic individuals picked up through screening for blood banks or other means. The individuals are diagnosed and then the diagnosis is confirmed via an HCV antibody test. Figure 1 shows the time and age variation in the reported number of HCV cases in Brazil.

In fact, the actual number of reported HCV infections is available only from 2000 onward. As we know from previous studies [25], HCV was introduced in Brazil in the later 1950s. We therefore constructed the number of reported with a sigmoidal decay backwards until 1932, as argued below. We used this artifice only to illustrate the model and these figures have little epidemiological significance, as argued below. We shall return to this point in the results section, where we explain this procedure in more detail.

**Figure 1 here**

**Estimating the total number of HCV infected individuals in Brazil**

Recall that $SINAN^*\{A,i\}$ is the number of individuals aged $A$ to $A$+1 at time $t_i$ who were notified to SINAN in the current year $i$, $(t_i\text{-}1,t_i]$. Now

$$SINAN^*\{A,i\} \approx \kappa_{A,i} I^{NN*}(A,i). \qquad (11)$$

This approximation is obtained by using equation (10) as

$$SINAN^*\{A,i\}=\int_A^{A+1} NOTIFICATION(a,(t_i-1,t_i])da.$$

As HCV infection is determined by taking an antibody test it is not possible to distinguish between individuals protected by maternal antibodies from HCV infected individuals. Hence we do not use the data for $A$=0 as it is unreliable, instead we take $SINAN^*\{0,i\} = 0$, for all $i$. Because only a very small number of individuals of age 0 are infected this does not cause significant error in the estimation.

From (7) and (11) we can write down the fundamental equation for estimating the incidence, for $A \geq 0$:

$$INC\{A,i\} = \frac{SINAN^*\{A,i\}}{\kappa_{A,i}} - \frac{SINAN^*\{A-1,i-1\}}{\kappa_{A-1,i-1}} \exp\left\{-\frac{1}{2}\left(\kappa_{A-1,i}+\kappa_{A,i}+\phi^{NN}_{A-1,i}+\phi^{NN}_{A,i}\right)\right\}, \qquad (12)$$

where $SINAN^*\{0,i\}$ and $SINAN^*\{\text{-}1,i\}$ are interpreted as zero for all $i$.

Note that, as observed in equation (12), the method consists of subtracting consecutive values of a diagonal of a matrix containing age in lines and time in columns. In some instances, however, it may happen that for certain ages and years the calculated incidence is negative. Our interpretation is that, for that particular age and time, the notified incidence was zero. When this happened in the actual calculation we assigned the value zero to the notification incidence.

Therefore, $I^{NN*}\{A,i\}$ can be calculated for each age and time reported as

$$I^{NN*}\{A,i\} = \sum_{j=0}^{A} INC\{A-j, i-j\} \exp\left\{-\frac{1}{2}\sum_{p=0}^{j-1}\left(\kappa_{A-1-p,i-p} + \kappa_{A-p,i-p} + \phi^{NN}_{A-1-p,i-p} + \phi^{NN}_{A-p,i-p}\right)\right\}. \quad (13)$$

Similarly, for $I^{N*}\{A,i\}$, we have:

$$I^{N*}\{A,i\} = \sum_{j=0}^{A} SINAN^{*}\{A-j, i-j\} \exp\left\{-\frac{1}{2}\sum_{p=0}^{j-1}\left(\phi^{N}_{A-1-p,i-p} + \phi^{N}_{A-p,i-p}\right)\right\}. \quad (14)$$

Figure 2 shows the calculation of $INC\{A,i\}$ using equation (12) with the SINAN data as shown in Figure 1.

**The Size of the Liver Transplantation Waiting List in Brazil**

It is known that a fraction of those individuals infected with HCV evolve to liver failure after many years of infection [26]. Let us denote those individuals diagnosed with liver failure of whose age in whole years is *A* at the end of calendar year *i*, time $t_i$ as $LF\{A,i\}$. These individuals have been necessarily diagnosed with HCV and, therefore, are a fraction of the notified infected

**Figure 2 here**

individuals $I^{N*}\{A,i\}$. It is assumed that individuals develop liver failure after a minimum time interval $\tau_{min}$, say 10 years. From equation (8) for $I^{N*}\{A,i\}$ we obtain the equation for $LF\{A,i\}$:

$$LF\{A,i\} = \sum_{\tau=\tau_{min}}^{A} \eta_{A-\tau}\, I^{N*}\{A-\tau, i-\tau\} \exp\left\{-\frac{1}{2}\left[\sum_{p=0}^{\tau-1}\left(\varphi^{N}_{A-1-p,i-p} + \varphi^{N}_{A-p,i-p}\right)\right]\right\},$$

(15)

where $\eta_{A-\tau}$ is a discretised function that decreases from $\tau = \tau_{min}$ up until $\tau = A$, representing the rate at which infected (and notified) individuals of age $A$-$\tau$ develop liver failure.

We know that liver damage (whether due to HCV or some other cause) is a progressive disease [27, 28] so the longer that an individual has been infected the more liver damage they will have sustained and the greater the chance of liver failure. Given a group of individuals currently all of age $A$ those that have been in the database longer are also more likely to have been infected for longer. Hence, $\eta_{A-\tau}$, the liver failure rate of those of current age $A$ who were notified to the database $\tau$ years ago should increase with $\tau$. Since early symptoms of liver disease precede complete failure it is reasonable to assume that there is a minimum gap between notification and liver failure.

Summing up over all ages we obtain the size of $LF\{i\}$, which is the total number of individuals with liver failure at time $t_i$:

$$LF\{i\} = \sum_{A_{min}}^{A_{max}} \sum_{\tau=\tau_{min}}^{A} \eta_{A-\tau}\, I^{N*}\{A-\tau, i-\tau\} \exp\left\{-\frac{1}{2}\left[\sum_{p=0}^{\tau-1}\left(\varphi^{N}_{A-1-p,i-p} + \varphi^{N}_{A-p,i-p}\right)\right]\right\}, \tag{16}$$

where $A_{\min}$ and $A_{\max}$ are minimum and maximum ages. Apart from those individuals who are transplanted (see below) $LF\{i\}$ corresponds to the Liver Transplantation Waiting List (LTWL).

Let us now rewrite equation (16) considering transplantation. Let $\psi(a,t)$ be the transplantation rate of individuals of aged $a \in [A, A+1)$ in calendar year $t \in (t_i\text{-}1, t_i]$. Then, equation (16) becomes

$$LWTL\{i\} = \sum_{A_{min}}^{A_{max}} \sum_{\tau=\tau_{min}}^{A} \eta_{A-\tau}\, I^{N*}\{A-\tau, i-\tau\} \exp\left\{-\frac{1}{2}\left[\sum_{p=0}^{\tau-1}\left(\varphi^{N}_{A-1-p,i-p} + \varphi^{N}_{A-p,i-p} + \psi^{N}_{A-1-p,i-p} + \psi^{N}_{A-p,i-p}\right)\right]\right\}. \tag{17}$$

The number of transplants in calendar year $i$ is then given by $TR\{i\}$ where

$$TR\{i\} = \sum_{A_{min}}^{A_{max}} \sum_{\tau=\tau_{min}}^{A} \psi_{A,i}\, \eta_{A-\tau}\, I^{N*}\{A-\tau, i-\tau\}$$
$$\exp\left\{-\frac{1}{2}\left[\sum_{p=0}^{\tau-1}\left(\varphi^{N}_{A-1-p,i-p} + \varphi^{N}_{A-p,i-p} + \psi^{N}_{A-1-p,i-p} + \psi^{N}_{A-p,i-p}\right)\right]\right\}. \quad (18)$$

We take for $\psi_{A,i}$ a suitably truncated bell-shaped discrete function [26] with a maximum at 45 years of age for all *i*.

**Results**

One of our objectives is to calculate equations (13) and (14) in order to obtain the estimated prevalence of notified and non-notified HCV infections which sum up to total prevalence. Unfortunately, the data available are restricted to the period between 2000 and 2012. In order to simulate a longer history of HCV infection in Brazil, we artificially constructed such a previous history by extrapolating backwards. First, we averaged the notified cases in the period between 2000 and 2012. Then, we fitted a sigmoidal-shaped curve representing the notified cases back for the period between 1932 and 2000. We did that for all ages such that the age distribution of notified cases was assumed fixed for all the extrapolated periods. We are well aware that HCV was probably introduced in Brazil in the 1950's and, therefore, this calculation is only an exercise to illustrate the method.

In a previous paper [16], this extrapolation was done differently. We assumed the disease to be in steady state until 1932. The results of this previous calculation are therefore different from the ones presented in this paper. We shall elaborate on this later. To begin with, Figure 3 shows a preliminary result on this direction. The continuous line is the total prevalence extrapolating the data as if in steady state [16]. The sigmoid dotted line is the total prevalence calculated assuming the artificially constructed notification as explained above.

**Figure 3 here**

Results of the numerical calculations are summarised in Table 2. In it we compare the prevalence in 2012 of HCV infected individuals who have been reported to SINAN until 2012 with the outcomes of the model. In Figure 4 we also compare the size of the Liver Transplantation Waiting List according to the official figures with the outcomes of the model.

Amaku et al. [16] assumed a stationary situation so time dependence was removed from the equations. A system of differential equations was used to describe the densities with respect to age of susceptibles, reported individuals, non-reported individuals and recovered individuals. One parameter was the disease reporting rate $\kappa$. They used two methods.

In the first method it was assumed that the age-dependent force of infection $\lambda(a)$ has a Gaussian shape with three scaling parameters. For a given value of $\kappa$ the force of infection was used in the differential equations and was parametrically fitted to the age-dependent SINAN incidence data. The value of $\kappa$ was then fitted heuristically to both the full age and time dependent SINAN data and the length of the LTWL. The fitted values of both $\lambda(a)$ and $\kappa$ were then used to find the total notified and non-notified HCV incidence data.

In the second method a different parametric function was fitted to the age-dependent SINAN incidence data. Given a value of $\kappa$ they next used the differential equations to model the incidence. Again the value of $\kappa$ was then fitted heuristically to both the full age and time dependent data and the length of the LTWL. The final fitted values of $\kappa$ and the SINAN age-dependent incidence data were used to find the total notified and non-notified HCV incidence data.

**Table 2. Summary of the Results**

| **RESULTS** | Current Method | First Method of [16] | Second Method of [16] |
|---|---|---|---|
| **Prevalence of Notified HCV Infections** | **163,902*** <br> **169,382**** | **- 240,120#** | **- 227,074#** |
| **Prevalence of Non-Notified HCV in Brazil** | **1,433,638*** <br> **1,446,771**** | **- 1,650,100#** | **- 1,632,300#** |

| | | | |
|---|---|---|---|
| **Total Prevalence of HCV in Brazil** | **1,597,540*<br>1,616,153**** | **-<br>1,890,220#** | **-<br>1,859,374#** |

*Using only the official SINAN period (2000-2012) assuming zero notification incidence for all years and ages from 2000 backwards until 1932.
** Calculated from real data (2000-2012) and extending the data backwards assuming a sigmoidal decay until 1932.
# Taking the average number of cases reported annually to SINAN between 2004 and 2012, a period in which a steady state could be assumed.

The corresponding results, called the first method and second method in Table 2, were obtained using the following procedure. First, we assumed that the infection was in steady state from 2004 to 2012 and averaged the reported incidence. This reported incidence was extrapolated backwards until 1932. It is therefore not surprising that the published numbers in [16] including the third and fourth columns of Table 2 are larger than the figures obtained in this paper. The difference represents up to a certain point the state of the infection prior to 2000 and from this point of view the results seem to be consistent with what was believed about the infection in Brazil.

From the results of the current method expressed in Table 2 it is possible to observe that the difference between taking into account the constructed data backwards until 1932 and the official SINAN period of 2000-2012, reflects the significant contribution of this period to both the SINAN and the total prevalence of HCV in Brazil. Note that the artificially constructed incidence will manifest itself for individuals older than 40 years.

Figure 4 shows the comparison between the actual size of the LTWL as in [17] Chaib et al. (2014) and the result of the application of equation (17). The parameter $\kappa$ was obtained in [16] by fitting the model to the LTWL. All other parameters were obtained independently of the LTWL. Figure 4 shows that using just this one fitted parameter the model accurately reproduces the whole LTWL time series. So we can assess the model as being reasonably accurate.

**Figure 4 here**

## Discussion

This paper is an attempt to provide a method to estimate the actual number of infected individuals (and other parameters related to transmission) of compulsory notifiable infectious diseases from the officially notified number of cases. Considering that, in the great majority of cases, the number of notified cases represents only a small but variable fraction of the total number of infected individuals, a reliable method of estimating the latter from the former can represent an important tool for public health policies. Notwithstanding the recognised importance of under-notification of most chronic infections, the tools to deal with this information gap proposed so far are varied and, to the best of our knowledge, there is currently no consensus about which is or are the most appropriate [3-8].

In a previous publication [16], a continuous time-dependent model for the estimation of the total number of HCV infected individuals in Brazil was proposed. In that paper, we assumed a steady state for the period between 2004 and 2012, and we concluded that the non-notified to notified ratio in the number of infections was about 7 to 1. The current work is an extension of that paper and we relaxed the steady state assumption. To do a calculation for individuals with age up to 80 years, we artificially extended the official notification database backwards from the year 2000 back to 1932. This artificially constructed database was intended only to illustrate the method. In addition, we discretised the variables time and age both because the notification database presents the number of cases per year and because the discrete model is easier to be implemented, both mathematically and computationally, than the continuous age and time corresponding model.

HCV is recently becoming virtually a 100%-curable disease due to antiviral treatments such as Ledipasvir/Acetonate/Sofosbuvir and others. So, there will be fewer and fewer individuals waiting for liver transplantation because of that. It is straightforward to modify the theoretical model to take account of this. If we have data on age, treatment and cure rates of individuals, let $\xi(a,t)$ denote the rate at which notified infectious individuals of age *a* are given treatment and cured at time *t*. Then in the continuous model (2) in the first partial differential equation for $S(a,t)$ there is an extra term

$$+\xi(a,t)I^N(a,t)$$

corresponding to infectious, notified, treated individuals who are cured and in the third partial differential equation of (2) for $I^N(a,t)$ the term

$$-\left(\mu(a,t)+\alpha^N(a,t)+\gamma^N(a,t)\right)I^N(a,t)$$

becomes

$$-\left(\mu(a,t)+\alpha^N(a,t)+\gamma^N(a,t)+\xi(a,t)\right)I^N(a,t),$$

so $\phi^N(a,t)$ becomes

$$\phi^N(a,t)=\mu(a,t)+\gamma^N(a,t)+\alpha^N(a,t)+\xi(a,t).$$

Thus it is straightforward to model antiviral treatment.

The method presented in this paper is applicable to any compulsory notifiable infectious disease provided that one has information about at least two end-points of the natural history of the disease of interest, or carrying out an alternative diagnostic test in a representative sample of the affected population. For instance, for the case of HCV, we used the number of notified cases and the size of the Liver Transplantation Waiting List. For other diseases, in which one has only the number of notified cases, an alternative to the Liver Transplantation Waiting List depends on the disease one is interested in. For instance, for the case of dengue in a sufficiently small region, an age-dependent seroprevalence profile of a properly designed sample of the population would be sufficient. For infections like HIV, in addition to the reported number of cases, a sample representing each group of risk should be used.

The method demonstrated to be accurate in retrieving the number of infected individuals for the case of HCV as it fits the Liver Transplant Waiting List data (see Figure 4) and the results are in good accordance with the previous estimations by Amaku et al. [16].

We have already said that the notification rate is the most important parameter in the model. This could be improved by various methods, for example public education about risk factors for HCV such as injecting drug use and new treatments, publicity campaigns, or screening programs, either of the general public or targeted high risk populations. Most important, however, would be a population-based seroprevalence study that could unequivocally determine individuals previously infected by HCV. The

ratio of notified individuals to seropositive ones would determine the actual value of notification rate ($\kappa$).

In spite of its accuracy and simplicity, the method here presented has some important limitations that are worthwhile mentioning. Firstly, the model is data-greedy in the sense that a long time series of notified cases is necessary for the calculations. Secondly, the model has a large number of parameters whose values are not known with any precision for the great majority of cases. For example, as the model deals with long time series, demographic parameters such as the natural mortality rate are crucial for the calculations.

Notwithstanding those limitations, the model has the advantage that it can predict quantities that can be iteratively used to improve it. For instance, for HCV the model allows the calculation of the proportion of individuals that have the infection for $\tau$ years, that is the age of infection. If this can be checked from information from patients (e.g., blood transfusion time), the model can be improved immediately. This is thoroughly explained in [16] Amaku et al. (2016).

**Conclusions**

We can conclude that the model proposed in this paper can be useful for estimation of the actual magnitude of endemic states of infectious diseases, particularly for those where the number of notified cases is only the tip of the iceberg. In addition, the method can be applied to other situations, such as the well-known under-reported incidence of criminality (for example rape), among others.

**List of abbreviations**

CND: Compulsory notifiable diseases
WHO: World Health Organization
IHR: International Health Regulations
HIV: human immunodeficiency virus
HCV: hepatitis C virus
LTWL: Liver Transplantation Waiting List
SIR: Susceptible-Infected-Removed

SINAN: Sistema de Informação de Agravos de Notificação (National Information System of Notifiable Diseases)

## Declarations

**Ethics approval and consent to participate.** This is a theoretical work based on secondary data in which no patients name has not been disclosed. No human subject has been recruited and therefore, there was no need of approval by any ethical committee.

**Consent for publication.** All authors agreed with the form and content of this manuscript as it is submitted.

**Availability of data and material.** All data used in this work are from a public database (SINAN) of the Brazilian Ministry of Health. This is publicly available through the internet. All the details of the deductions and calculations are presented in the manuscript.

**Competing interests.** None.

**Funds** This work was partially funded by LIM01-HCFMUSP, CNPq, Brazilian Ministry of Health (Grant TED 27/2015) and FAPESP. DG is grateful to the Leverhulme Trust for support from a Leverhulme Research Fellowship (RF-2015-88) and the British Council, Malaysia for funding from the Dengue Tech Challenge (Application Reference DTC 16022). EM and DG are grateful to the Science Without Borders Program for a Special Visiting Fellowship (CNPq grant 30098/2014-7).

**Authors' contributions.** MA, FABC and EM designed the model. DG, MA, FABC, MNB, EM and LFL developed the deductions and calculations. EC and EM calculated the liver transplantations waiting list part of the model. EM, FACB and MNB wrote the paper.

## Appendix

In this Appendix, we deduce the equation (7) from the main text. Let us define the function $I^{NN}(a+x,t+x)$, which is a function that expresses the evolution of a cohort. Then

$$\frac{d}{dx}[I^{NN}(a+x,t+x)] = \lambda_d(a+x,t+x)\,S(a+x,t+x) \\ -[\kappa_d(a+x,t+x)+\phi_d^{NN}(a+x,t+x)]I^{NN}(a+x,t+x)\,, \tag{A1}$$

where

$$\phi_d^{NN}(a+x,t+x) = \mu_d(a+x,t+x)+\gamma_d^{NN}(a+x,t+x)+\alpha_d^{NN}(a+x,t+x).$$

Multiplying both sides by $\exp\left[\int_0^x\left(\kappa_d(a+z,t+z)+\phi_d^{NN}(a+z,t+z)\right)dz\right]$, we have

$$\frac{d}{dx}\left[\exp\left[\int_0^x\left(\kappa_d(a+z,t+z)+\phi_d^{NN}(a+z,t+z)\right)dz\right]I^{NN}(a+x,t+x)\right] = \\ \lambda_d(a+x,t+x)\,S(a+x,t+x)\exp\left[\int_0^x\left(\kappa_d(a+z,t+z)+\phi_d^{NN}(a+z,t+z)\right)dz\right]. \tag{A2}$$

So integrating we deduce that

$$I^{NN}(a,t) = I^{NN}(a-1,t-1) \tag{A3}$$
$$\exp\left[-\int_0^1\{\kappa_d(a-1+z,t-1+z)+\phi_d^{NN}(a-1+z,t-1+z)\}dz\right]$$
$$+\int_0^1\lambda_d(a-1+x,t-1+x)S(a-1+x,t-1+x)$$
$$\exp\left[-\int_x^1\{\kappa_d(a-1+z,t-1+z)+\phi_d^{NN}(a-1+z,t-1+z)\}dz\right]dx.$$

The first term corresponds to non-notified individuals ages *a*-1 at time *t*-1 who remain infectious and non-notified at time *t* (when their age is *a*). The second term which we denote

$$INCIDENCE(a,(t\text{-}1,t])$$

is the density with respect to age *a* of the incidence of HCV in the cohort of individuals born at time *t*-*a* which occurs in the time interval (*t*-1,*t*] and is still infectious and not notified at time *t*.

Now, $I^{NN*}\{A,i\}$, the absolute number of infectious non-notified individuals of age in the interval $[A,A+1)$ at time $t_i$,

$$= \int_A^{A+1} I^{NN}(a,t_i)da, \tag{A4}$$

$$\approx I^{NN}\left(A+\frac{1}{2},t_i\right), \tag{A5}$$

taking the midpoint as an approximation.

Now from (A3) and (A4)

$$I^{NN*}\{A,i\} = \int_A^{A+1} I^{NN}(a-1,t_i-1)$$

$$\exp\left[-\int_0^1 \{\kappa_d(a-1+z,t_i-1+z)+\phi_d^{NN}(a-1+z,t_i-1+z)\}dz\right]da$$

$$+\int_A^{A+1} INCIDENCE(a,(t_i-1,t_i])da, \tag{A6}$$

where for $a \leq 0, I^{NN}(a,t)$ is interpreted as zero. The last term in (A6), which we shall denote $INC\{A,i\}$, represents the incidence between times $t_i$-1 and $t_i$ of HCV that is still infectious and not notified at time $t_i$, in the cohort born between times $t_i$-$A$-1 and $t_i$-$A$. In the first term in (A6) again for the $a$-integration we take $a$=$A$+$\frac{1}{2}$ as an approximation, as the integration interval has length one.

$$I^{NN*}\{A,i\} \approx I^{NN}\left(A-\frac{1}{2},t_i-1\right)$$

$$\exp\left[-\int_0^1 \left\{\kappa_d\left(A-\frac{1}{2}+z,t_i-1+z\right)+\phi_d^{NN}\left(A-\frac{1}{2}+z,t_i-1+z\right)\right\}dz\right]$$

$$+INC\{A,i\}.$$

$$= I^{NN}\left(A-\frac{1}{2},t_i-1\right)$$

$$\exp\left[-\int_0^1 \left\{\kappa_d\left(A-\frac{1}{2}+z,t_i\right)+\phi_d^{NN}\left(A-\frac{1}{2}+z,t_i\right)\right\}dz\right]+INC\{A,i\},$$

as $\kappa_d\left(A-\frac{1}{2}+z,t\right)$ and $\phi_d^{NN}\left(A-\frac{1}{2}+z,t\right)$ are the same for $t \in (t_i-1,t_i]$.

$$\approx I^{NN*}\{A-1,i-1\}\exp\left[-\frac{1}{2}\left(\kappa_{A-1,i}+\kappa_{A,i}+\phi^{NN}{}_{A-1,i}+\phi^{NN}{}_{A,i}\right)\right]$$

$$+INC\{A,i\},$$

because

(i) Noting that year $i$-1ends at time $t_i$-1 we have

$$I^{NN}\left(A-\tfrac{1}{2},t_i-1\right) \approx I^{NN*}\{A-1,i-1\}, \quad \text{by (A5).}$$

(ii) for $z \in \left[0, \frac{1}{2}\right)$, $\kappa_d\left(A - \frac{1}{2} + z, t_i\right) = \kappa_{A-1,i}$ and for $z \in \left[\frac{1}{2}, 1\right]$, $\kappa_d\left(A - \frac{1}{2} + z, t_i\right) = \kappa_{A,i}$.

**References**


1. Roush S, Birkhead G, Koo D, Cobb A and Fleming, D. Mandatory reporting of diseases and conditions by health care professionals and laboratories. *JAMA* 1999;282:164–170.

2. MMWR - Summary of notifiable diseases, United States, 1998. *MMWR Morb Mortal Wkly Rep* 1999;47:ii–92.

3. Doyle, TJ, Glynn, MK and Groseclose, SL. Completeness of notifiable infectious disease reporting in the United States. *Am J Epidemiol* 2002;155:866–874.

4. Gibbons et al. *BMC Public Health* 2014, 14:147 Page 2 of 17. Available at: http://www.biomedcentral.com/1471-2458/14/147. *Accessed at 11 October 2016.*

5. Keramarou, M and Evans, MRE. Completeness of infectious disease notification in the United Kingdom: A systematic review. *J Infection* 2012;64:555-564.

6. Rowe, SL and Cowie, BC. Using data linkage to improve the completeness of Aboriginal and Torres Strait Islander status in communicable disease notifications in Victoria. *Aust NZ J Public Health* 2016;40:148-53; doi: 10.1111/1753-6405.12434

7. Gibney, KB, Cheng, AC, Hall, R and Leder, K. An overview of the epidemiology of notifiable infectious diseases in Australia, 1991–2011. *Epidemiol Infect* 2016;144(15); 3263-3277. doi:10.1017/S0950268816001072.

8. Serra I, García V, Pizarro A, Luzoro A, Cavada G and López J. A universal method to correct -reporting of communicable diseases. Real incidence of hydatidosis in Chile, 1985-1994. *Rev Med Chi* 1999 Apr;127(4):485-492.

9. Ximenes R, Amaku M, Lopez LF, Coutinho FAB, Burattini MN, Greenhalgh D, Wilder-Smith A, Struchiner CJ and Massad E. The risk of dengue for non-immune foreign visitors to the 2016 summer Olympic games in Rio de Janeiro, Brazil, 2016; *BMC Inf Dis*, 16:Article No 186.

10. Konowitz PM, Petrossian GA and Rose DN. The under-reporting of disease and physician's knowledge of reporting requirements. *Pub Health Rep* 1984; Jan-Feb; 99(1):31-35.

11. Rosenberg ML, Gangarosa EJ, Pollard RA, Wallace M, Brolnitsky O and Marr JS. *Shigella* surveillance in the United States, 1975. *J Infect Dis* 1977; 136:458-460.

12. Brabazon, ED, O'Farrell A, Murray CA, Carton MW and Finnegan P. Under-reporting of notifiable infectious disease hospitalization in a health-board region in Ireland: room for improvement?, *Epidemiol Inf* 2008; Feb; 136(2):241-247.

13. Thacker SB, Choi K and Brachman PS. The surveillance of infectious diseases. *JAMA* 1983;249:1181–1185.

14. Schiffman EK, McLaughlin C, Ray JAE, Kemperman MM, Hinckley AF, Friedlander HG and Neitzel DF. Under-reporting of Lyme and other Tick-Borne diseases in residents of a high-incidence county, Minnesota, 2009, *Zoonoses Pub Health*, 2016; doi:10.111/zph.12291.

15. Mann JM. Health and human rights: broadening the agenda for health professionals. *Health Hum Rights* 1996;2(1):1–5.

16. Amaku, M, Burattini, MN, Coutinho, FAB, Lopez, LF, Mesquita, F, Naveira, MCM, Pereira, GFM, Santos, ME and Massad, E. Estimating the size of the HCV infection prevalence: A modeling approach using the incidence of cases reported to an official notification system. *Bull Math Biol* 2016;78:970–990. doi: 10.1007/s11538-016-0170-4.

17. Chaib E, Massad E, Varone BB, Bordini AL, Galvão FHF, Crescenzi A, Filho AB and D'Albuquerque LA. The impact of the introduction of MELD on the dynamics of the Liver Transplantation Waiting List in São Paulo, Brazil. *J Transplant* 2014; 2014:219789. doi: 10.1155/2014/219789.

18. MHB **-** Inf. Epidemiol. Sus v.9 n.1 Brasília mar. 2000. Available at: http://dx.doi.org/10.5123/S0104-16732000000100006. *Accessed at 10 June 2016.*

19. Trucco, E. Mathematical models for cellular systems: The Von Foerster Equation. *Bull Math Biophys* 1965;27:285-304.

20. Lopez LF, Amaku M, Coutinho FA, Quam M, Burattini MN, Struchiner CJ, Wilder-Smith A and Massad E. Modeling importations and exportations of infectious diseases via travelers. *Bull Math Biol* 2016 Feb;78(2):185-209. doi: 10.1007/s11538-015-0135-z.

21. Roberts EA and Yeung L. Maternal-infant transmission of hepatitis C virus, *Hepatology*, 2002; 36(S1):S106-S113.

22. WHO Factsheet Hepatitis C. Available at: http://www.who.int/mediacentre/factsheets/fs164/en/. *Accessed at 10 October 2016.*

23. SINAN, Sistema de Informação de Agravos de Notificação. Available at: http://portalsinan.saude.gov.br/hepatites-virais. *Accessed at 10 December 2015.*

24. Aguiar M, Rocha F, Pessanha JEM, Mateus L and Stollenwerk N. Carnival or football, is there a real risk for acquiring dengue fever during holidays seasons*? Sci Rep* 2015; 5:8462.

25. Romano CM, de Carvalho-Mello IM, Jamal LF, de Melo FL, Iamarino A, Motoki M, Pinho JR, Holmes EC, de Andrade Zanotto PM and the VGDN Consortium. Social networks shape the transmission dynamics of hepatitis C virus. *PLoS One* 2010 Jun 23;5(6):e11170. doi: 10.1371/journal.pone.0011170.

26. Chaib E and Massad E. Liver transplantation: waiting list dynamics in the state of São Paulo, Brazil. *Transplant Proc* 2005 Dec;37(10):4329-4330.

27. American Liver Foundation. The progression of Liver disease. Available at: www.liverfoundation.org/abouttheliver/info/progression *Accessed at 24 August 2017*.

28. NHS Choices. Hepatitis C. Available at: www.nhs.org/conditions/Hepatitis-C/Pages/Introduction.aspx *Accessed at 24 August 2017*.

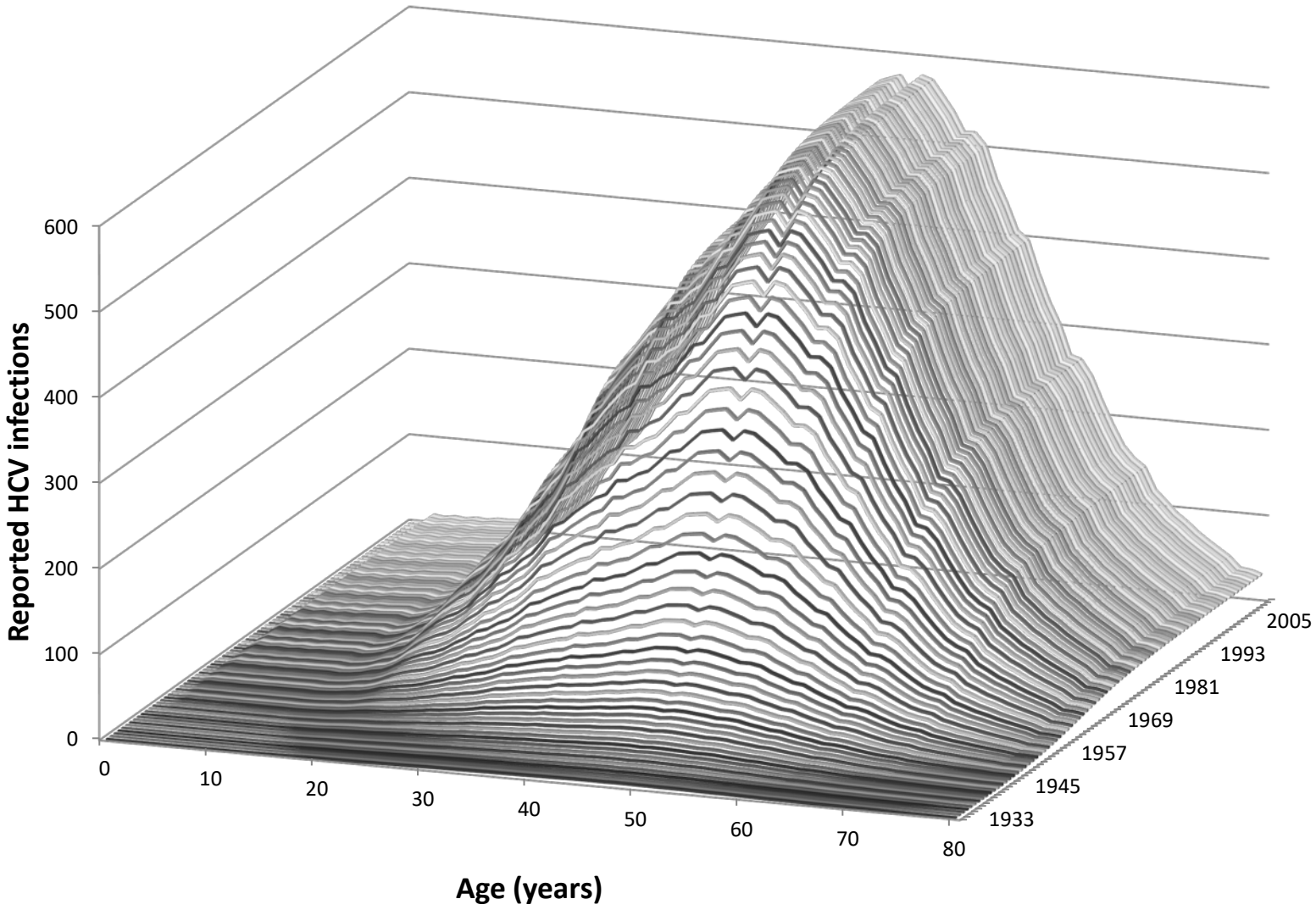
## Legends to the Figures:

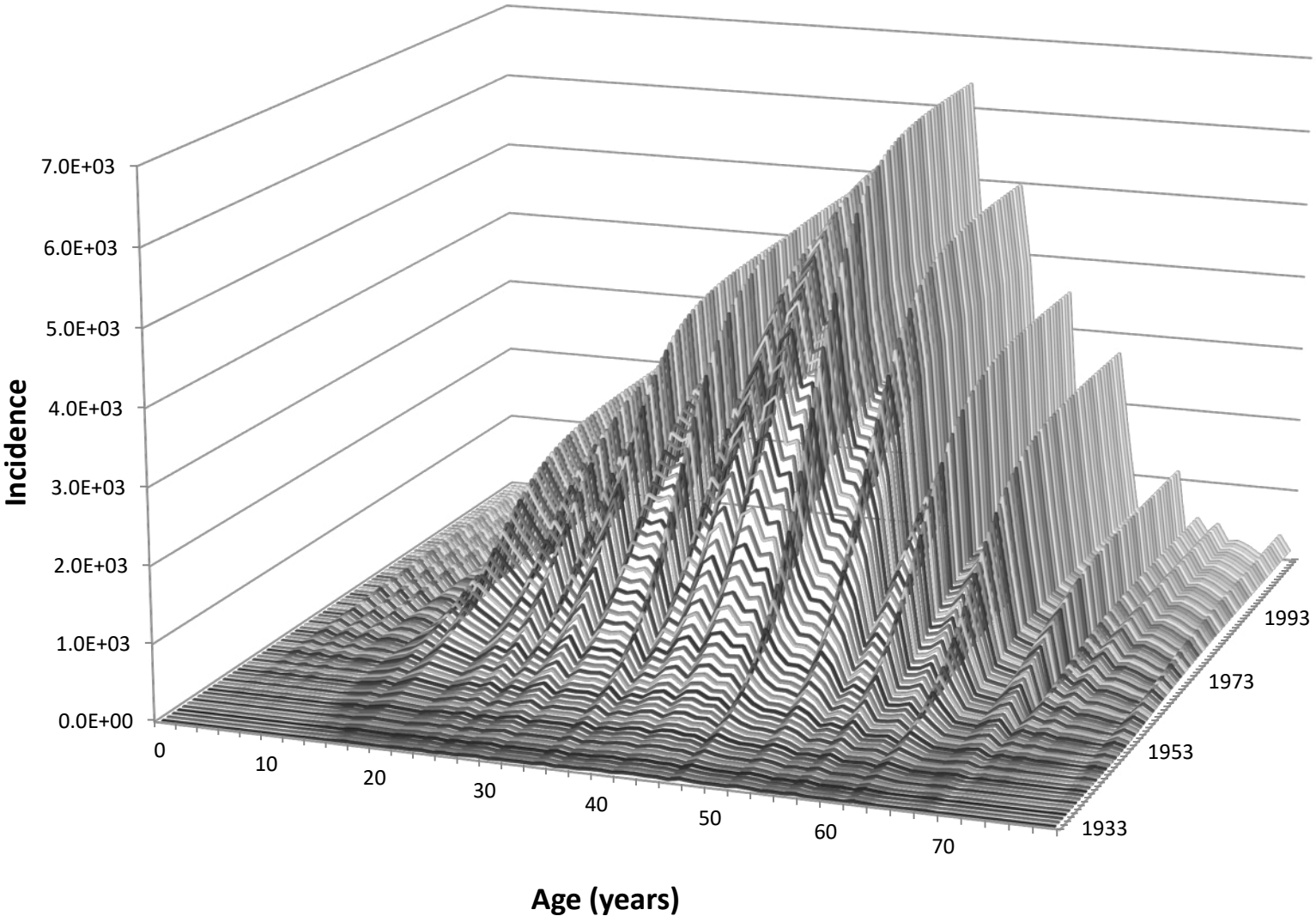
**Figure 1. Time and Age variation of the reported number of HCV infections in Brazil, artificially constructed by extrapolating backwards until 1932.**

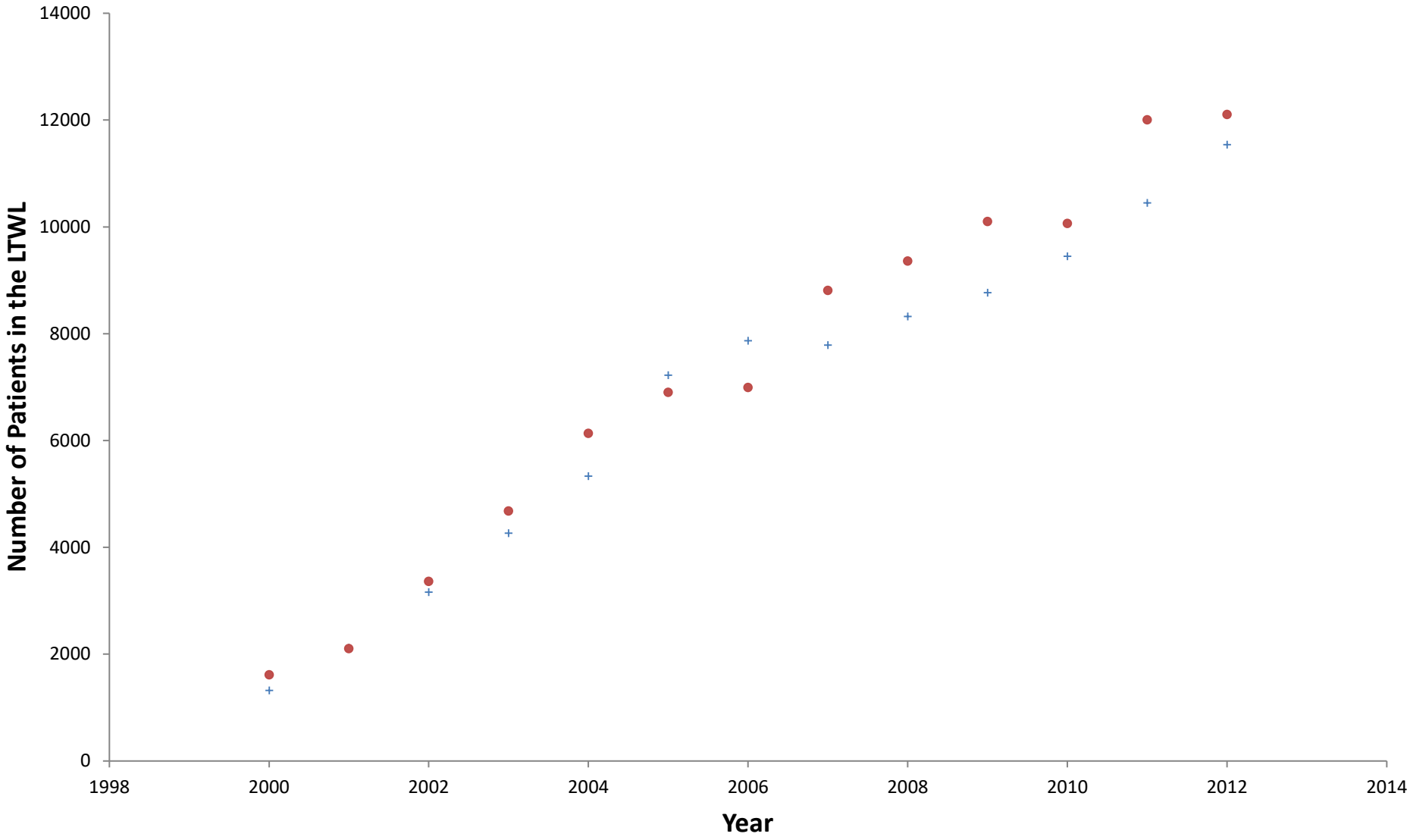
**Figure 2. Calculation of $INC\{A, i\}$ from the SINAN data as shown in Figure 1.**

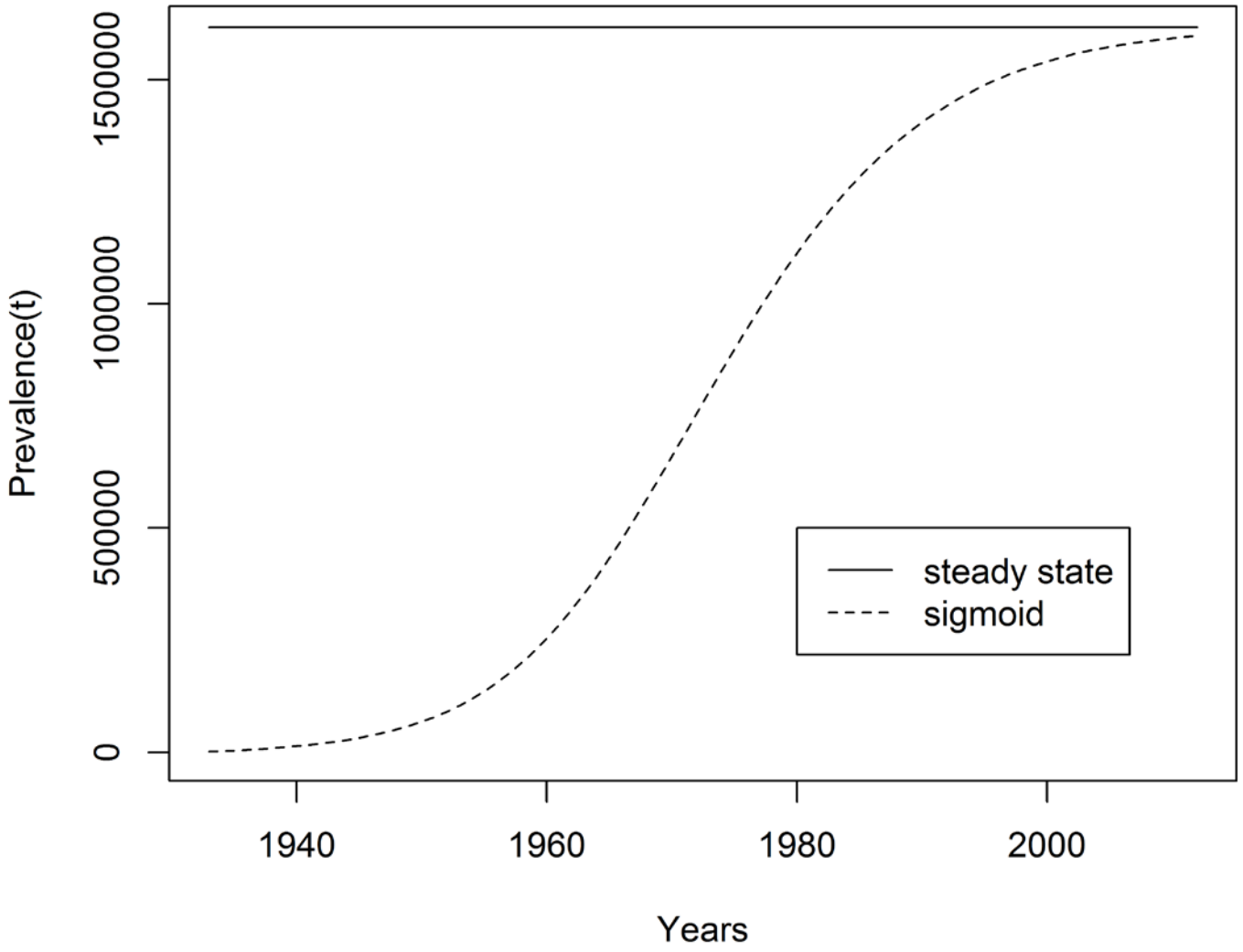
**Figure 3. Comparison of the total prevalence calculated according to Amaku et al. [16] (continuous line) and assuming the notification as a sigmoidal extrapolation (dotted line).**

**Figure 4. Comparison between the empirical data on the size of the LTWL (crosses) as in Chaib et al. [17] and the result of the application of equation (17) (dots).**


Reported HCV infections
600
500
400
300
200
100
0
0
10
20
30
40
50
60
70
80
Age (years)
1933
1945
1957
1969
1981
1993
2005


Incidence
7.0E+03
6.0E+03
5.0E+03
4.0E+03
3.0E+03
2.0E+03
1.0E+03
0.0E+00
0
10
20
30
40
50
60
70
Age (years)
1933
1953
1973
1993


Number of Patients in the LTWL
14000
12000
10000
8000
6000
4000
2000
0
1998
2000
2002
2004
2006
2008
2010
2012
2014
Year


Prevalence(t)
0
500000
1000000
1500000
1940
1960
1980
2000
Years
steady state
sigmoid